Geometric Stratification of Accounting Data

Patricia Gunning *
Jane Mary Horgan **
William Yancey ***

Abstract:

We suggest a new procedure for defining the boundaries of the strata in highly skewed populations, usual in auditing, which is much easier to use than the commonly used cumulative root frequency method of Dalenius and Hodges (1957, 1959). We implement it on two audit populations, one a population of debtors in an Irish firm, and the other a population of sales and use tax liabilities in the US. Our results show that the new method compares favourably with the cumulative root frequency method in terms of the accuracy of the estimates.

Keywords: Accounting data, statistical auditing, efficiency, sampling, stratification.

1. Introduction

One of the objectives of an audit is to estimate the total amount of error in a population of interest: for example, the amount of overstated debts in the set of transactions of a firm, or underpaid taxes in the set of transactions to which sales and use taxes apply. For large businesses, it is too costly and time consuming to examine all transactions in the population: instead a sample is selected and subjected to detailed testing in order to estimate the parameters of the whole population.

The distribution of accounting populations is usually highly skewed with long tails to the right. Therefore a stratification design is often used, where the population is divided into several strata and separate random samples are drawn from each stratum. The objective is to decide on a stratification design to minimise the variance of the resulting estimates.

Dalenius (1950) derived equations for determining boundaries for the strata so that the variance of the resulting estimate is minimised. He pointed out that these equations are troublesome to solve because of dependencies among the components, and suggested a method of approximation (Dalenius and Hodges, 1957), called the cumulative root frequency method, which is still today the most commonly used method of constructing stratum boundaries. This method involves first dividing the sorted frame into a fairly large number of classes, obtaining the root of the frequency in each class, accumulating the root of the frequencies, and obtaining strata so that there are equal intervals on the cumulative root frequencies. Though commonly used in practice, the cumulative root frequency method has some arbitrariness which makes it difficult to implement; the final stratum boundaries depend

* Researcher, School of Computing, Dublin City University, Dublin, Ireland.
Correo electrónico: pgunning@computing.dcu.ie
** Senior Lecturer, School of Computing, Dublin City University, Dublin, Ireland.
Correo electrónico: jhorgan@computing.dcu
*** Independent Consultant, Dallas, Texas.
Correo electrónico: will@willyancey.com

on the initial choice of the number of classes and there is no theory that gives the best number of classes (Hedlin 2000). In this paper, we suggest a new method of stratification, suitable for use with positively skewed populations, which is simpler than the solution of Dalenius and Hodges, and overcomes the arbitrariness. We examine its performance when applied to two real accounting populations, one a population of Irish debtors, and the other a population of US sales-and-use taxes.

The paper is structured as follows. We begin in Section 2 by overviewing stratification in statistical auditing, and go on in Section 3 to outline the procedures for constructing strata. In Section 4 the new procedure is implemented on two real accounting populations, and its efficiency is examined in Section 5. We conclude with a summary of our results.

2. Stratification in Auditing

Suppose there are N transactions in a population. Let $X_1, X_2, ...,X_N$ be the amounts stated by the debtor or taxpayer, and $Y_1, Y_2,....,Y_N$ be the true amounts to be ascertained by the auditor. The objective is to estimate the total true amount, T_Y, and compare it with the stated amount T_X. To do this a sample of size n is chosen and used to provide an estimate of T_Y.

When a stratification design is used, the transactions are ordered in increasing size of X, and divided into L mutually exclusive strata. The strata are intervals determined by their endpoints $k_0, k_1, ... ,k_L$. If $N_1, N_2,..., N_L$ represent the number of items within each stratum then the total population is given by $N=\Sigma_h N_h$. If $n_1,n_2,....,n_L$ represent the number of elements to be selected from each stratum then the total sample size is given by $n=\Sigma_h n_h$

One general strategy in choosing stratum boundaries is to choose them in such a way that the variance of the resulting estimate is minimised. For example, in the case of the stratified unbiased estimate of the total:

$$\hat{T}_{y,st} = N \sum_{h=1}^{L} \frac{N_h}{N} \overline{y_h}$$

where $\overline{y}_h$ is the mean of the sample elements in the hth stratum; we choose a design in order to minimise its variance:

$$V(\hat{T}_{y,st}) = N^2 \sum_{h=1}^{L} \left(\frac{N_h}{N} \right)^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2}{n_h}$$

where S_h is the standard deviation of all elements in stratum h, i.e.

$$S_h = \sqrt{\frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \overline{Y}_h)^2}$$

Here $\overline{Y}_h$ is the mean of the full set of elements $(Y_{hi}, i = 1, 2, \ldots. N_h)$ in stratum h.

Stratified sampling designs are distinguished by:

1. the number L of strata used;

2. how the sample is allocated among the strata;

3. the construction of the stratum boundaries.

As far as (1), L the number of strata is concerned, Cochran (1977) developed a model representing the approximate reduction in the variance gained over simple random sampling by stratification, and deduced that there is little to be gained from having more than five strata.

When it comes to (2), allocating the sample elements among strata, it can be shown that the variance of the stratified mean is minimised when

$$n_h = \left(\frac{N_h S_h}{\sum_{l=1}^{L} N_l S_l} \right) \cdot n$$

This method of sample allocation is detailed in Cochran (1977), and is referred to as Neyman allocation.

We look at (3), the construction of stratum boundaries, in the next section.

3. Construction of the Stratum Boundaries

Before detailing the new method, we summarise the commonly used cumulative root frequency method.

3.1. The Cumulative Root Frequency Method

The cumulative root frequency procedure of stratum construction is carried out for the division of a population into L strata as follows:

1. Arrange the stratification variable X in ascending order;
2. Group the X into a number of classes, J;
3. Determine the frequency in each class f_i $(i=1, 2, \ldots J)$;
4. Determine the square root of the frequencies in each class;
5. Cumulate the square root of the frequencies $\sum_{i=1}^{J} \sqrt{f_i}$
6. Divide the sum of the square root of the frequencies by the number of strata:

$$Q = \frac{1}{L} \sum_{i=1}^{J} \sqrt{f_i}$$

7. Take the upper boundaries of each stratum to be the X values corresponding to

$$Q,\ 2Q,\ 3Q, \ldots\ldots (L-1)Q,\ LQ.$$

Notice that these are in an arithmetic progression.

To illustrate the implementation of the cumulative root frequency method of stratum construction, we use a population of Irish debtors' stated liabilities consisting of 3,369 items ranging between €40 and €28,000. The data are divided into 30 initial classes, and given in Table 1 below; observe how the distribution is highly skewed, the low values *X* have a high incidence of occurrence, and the incidence decreases as the *X* values increase.

Table 1
Calculation of the Stratum Boundaries by the Cumulative Root Frequency Rule.

X-value	f_i	$\sqrt{f_i}$	Cum $\sqrt{f_i}$
40 - 972	2755	52.49	52.49
972 - 1904	258	16.06	68.55
1904 - 2836	137	11.70	80.26
2836 - 3768	60	7.75	88.00
3768 - 4700	42	6.48	94.48
4700 - 5632	33	5.74	100.23
5632 - 6564	18	4.24	104.47
6564 - 7496	14	3.74	108.21
7496 - 8428	11	3.32	111.53
8428 - 9360	9	3	114.53
9360 - 10292	6	2.45	116.98
10292 - 11224	2	1.41	118.39
11224 - 12156	6	2.45	120.84
12156 - 13088	2	1.41	122.25
13088 - 14020	4	2	124.25
14020 - 14952	0	0	124.25
14952 - 15884	4	2	126.25
15884 - 16816	0	0	126.25
16816 - 17748	0	0	126.25
17748 - 18680	1	1	127.25
18680 - 19612	2	1.41	128.67
19612 - 20544	0	0	128.67
20544 - 21476	1	1	129.67
21476 - 22408	0	0	129.67
22408 - 23340	1	1	130.67
23340 - 24272	0	0	130.67
24272 - 25204	1	1	131.67
25204 - 26136	0	0	131.67
26136 - 27068	1	1	132.67
27068 - 28000	1	1	133.67

To stratify the population into three strata, we calculate Q = 133.67/3 = 44.56. Then we take the upper boundaries of each stratum to be the X-values corresponding to 44.56, 89.11 and 133.67. The nearest available breaks are 972, 3,768 and 28,000, giving corresponding strata in the range:

40-972; 973-3,768; 3,769-28,000

This example illustrates not only the tedious nature of the calculations necessary to obtain the boundaries, but how the choice of the initial class intervals ultimately determines the boundary values. The final stratum boundaries depend on the choice of the number of classes. Had we chosen a different number of initial classes in Table 1, we would have obtained different boundaries.

3.2. The Geometric Method

The new method we are proposing to stratify skewed populations is based on an observation by Cochran (1961) that when the optimum boundaries of Dalenius (1950) are achieved, the coefficients of variation ($CV_h = S_{hx}/\overline{X}_h$) are often found to be approximately the same in all strata. i.e

$$\frac{S_{1x}}{\overline{X}_1} = \frac{S_{2x}}{\overline{X}_2} = ... = \frac{S_{Lx}}{\overline{X}_L}$$

Here S_{hx} is the standard deviation of the stated amounts in stratum *h*, and $\overline{X}_h$ is the mean.

Lavallée and Hidiroglou (1988) noted that the equality of coefficients is often asked by users of survey data. Gunning and Horgan (2004) have derived a simple algorithm for setting the coefficients of variation approximately equal in each stratum. They show that, if it can be assumed that the data are approximately uniformly distributed within each stratum, near-equal coefficients of variation may be achieved by constructing the stratum boundaries using the geometric progression.

This geometric stratification is implemented for dividing a population into *L* strata as follows:

1. Arrange the stratification variable *X* in ascending order;
2. Take the minimum value as the first term, and the maximum value as the last term of the geometric series with L+1 terms;
3. Calculate the common ratio: $r = (max/min)^{1/L}$;
4.Take the boundaries of each stratum to be the *X* values corresponding to the terms in the geometric progression with this common ratio:

$$Minimum\ k_0 = a,\ ar,\ ar^2 \ldots..\ ar^L = maximum\ k_L.$$

A full proof of the validity of this result is given in Gunning and Horgan (2004).

To illustrate its simplicity we apply it to stratify the data provided in Table 1 into three strata, i.e.

$$L=3,\quad k_0=40,.....,k_3=28{,}000:$$

thus $r = (28{,}000/40)^{1/3} = 700^{1/3} = 8.88$, and $k_h=40*8.88^h$ $(h=0,1,2,3)$.

Therefore, the strata form the ranges

40-354; 355-3,152; 3,153-28,000.

This is clearly a much simpler method of obtaining stratum breaks than the cumulative root frequency method. It is suitable for positively skewed populations, common in accounting data, where the low values of the variable have a high incidence of occurrence, and the incidence decreases as the variable values increase. It is therefore appropriate to take small intervals at the beginning and large intervals at the end. This is what happens with a geometric series of constant ratio greater than one. In the lower range of the variable, the strata are narrow so that an assumption of a uniform distribution is not unreasonable. As the

value of the variable increases, the stratum width increases geometrically. This coincides with the decreased rate of change of the incidence of the positively skewed variable, so here also the assumption of uniformity is reasonable.

We expect the best results, in terms of equality of coefficients of variation, when the distribution is highly positively skewed and the upper part contains a small percentage of the total frequency, and a large percentage of the monetary amount. Again this can be said to be an apt description of accounting data.

In the next section, we examine how successful this algorithm is in equalising the coefficients of variation in the strata.

4. Comparison of the Methods of Stratum Construction

To test our algorithm, we implement it on two specific accounting populations, both of which are positively skewed.

4.1. The Accounting Data

Our first population (Population 1) is that already given in Table 1 and consists of debtors' stated liabilities in an Irish firm; it is further detailed in Horgan (2003). Our second population (Population 2) is a population of stated liabilities of taxpayers in a US firm. The audit often consists of a take-all stratum of very large items that are audited on a 100% basis, and a take-none stratum of very small items that are not audited at all. In Population 1, the take-all stratum consisted of all items over €28,000, and the take-none stratum consisted of all items under €40. In Population 2, the take-all stratum consisted of all items over $50,000, and the take-none stratum consisted of all items under $100. In each population, the remainder of the population is sampled. The trimmed populations that are sampled are summarised in Table 2

Table 2
Summary Statistics

Population	N	Range	Skewness	Mean	Variance
Population 1(€)	3, 369	40-28,000	6.44	828	3,511,827
Population 2($)	249,106	100-50,000	4.79	2,258	24,352,340

It is interesting to note from Table 2 that, although the two populations represent different types of accounting situations in different countries, they are both highly positively skewed. The US population contains a substantially greater number of items, and its skewness coefficient is smaller.

4.2. The Stratum Boundaries

The two populations described in Table 2 were divided into $L = 3, 4$ and 5 strata using the cumulative root frequency method (cumroot), and the new geometric method, and the coefficient of variation was calculated in each stratum. Tables 3, 4 and 5 give the results for $L = 3, 4$ and 5 strata respectively.

Table 3
Strata Construction: 3 Strata

Population	Method		1	2	3
1	Geometric	Range	40-354	355-3152	3153-28000
		% of Pop	56%	38%	6%
		CV_h	.71	.68	.64
	Cumroot	Range	40-558	559-2236	2237-28000
		% of Pop.	69%	22%	9%
		CV_h	.70	.42	.76
2	Geometric	Range	100-793	794-6298	6299-5000
		% of Pop.	55%	36%	9%
		CV_h	.82	.62	.61
	Cumroot	Range	100-1498	1499-4495	4496-50000
		% of Pop.	72%	17%	11%
		CV_h	.73	.35	.68

Table 4:
Strata Construction: 4 Strata

Population	Method		1	2	3	4
1	Geometric	Range	40-205	206-1057	1058-5443	5444-28000
		% of Pop	42%	41%	14%	3%
		CV_h	.45	.44	.48	.50
	Cumroot	Range	40-558	559-1117	1118-2795	2796-28000
		% of Pop.	69%	14%	10%	7%
		CV_h	.70	.19	.27	.69
2	Geometric	Range	100-472	473-2235	2236-10572	10573-50000
		% of Pop.	.40	.39%	16%	5%
		CV_h	45	.43	.47	.45
	Cumroot	Range	100-999	999-2497	2498-9991	9992-50000
		% of Pop.	53%	31%	12%	4%
		CV_h	.64	.27	.42	.48

Table 5
Strata Construction: 5 Strata

Population	Method		1	2	3	4	5
1	Geometric	Range	40-147	148-549	550-2037	2038-7552	7553-28000
		% of Pop	31%	37%	22%	8%	2%
		CV_h	.37	.38	.40	.37	.41
	Cumroot	Range	40-279	280-838	839-1677	1678-4193	4194-28000
		% of Pop.	49%	30%	10	7%	4%
		CV_h	.52	.30	.20	.25	.57
2	Geometric	Range	100-346	347-1200	1201-4161	4162-14,425	14,426-50,000
		% of Pop.	.33%	.34%	.21%	.9%	.3%
		CV_h	.35	.35	.37	.36	.37
	Cumroot	Range	100-998	999-1997	198-3996	3997-12489	12490-50000
		% of Pop.	.61%	.16%	.10%	.9%	.4%
		CV_h	.64	.20	.20	.33	.41

A cursory examination of the coefficients of variation in Tables 3, 4, and 5 suggests that the geometric method is more successful than the cumulative root frequency method in obtaining near-equal CV_h : the CV_h differ substantially more from each other when the cumulative root frequency method is used to make the breaks than when the geometric method is used.

A more detailed analysis of the variability of the CV_h between strata is given in Table 6, where the standard deviation of the CV_h is calculated for each design.

Table 6
The Variability of the CV_h for Each Design

Strata	Method	Population 1	Population 2
3	Geometric	.035	.119
	Cumroot	.181	.201
4	Geometric	.028	.016
	Cumroot	.271	.153
5	Geometric	.018	.010
	Cumroot	.166	.182

We see from Table 6 that the standard deviations of the CV_h are substantially lower with the geometric method of stratum construction than with the cumulative root method. The new algorithm appears to be successful in breaking the strata in a way that the CV_h are close in value. Those obtained with the cumulative root frequency method are substantially more variable. On the basis of the observation of Cochran (1961) this suggests that the variance of the mean is close to minimal with geometric stratification. We investigate this in the next section.

5. The Efficiency of Geometric Stratification

The geometric method is compared with the cumulative root frequency in terms of the relative efficiency (*Eff*), the ratio of the variances of the means obtained when the strata are constructed using each method.

$$Eff = \frac{V_{geom}(\bar{x}st)}{V_{cum}(\bar{x}st)}$$

In each case the sample elements are allocated optimally among the strata . In sample size planning, the relative efficiency represents the proportionate increase or decrease in the sample size with the cumulative root frequency method to obtain the same precision as that of the geometric method.

The variance calculations are based on *X*, the stated values of the debts (Population 1), and the stated taxpayers' liabilities (Population 2). Since *X* is highly correlated with the true values *Y*, we can assume that the relative efficiency *Eff*, given above, is a reasonable approximation of the relative efficiency of *Y*.

The relative efficiencies are given in Tables 7 and 8 for Population 1 and Population 2 respectively. The sample size *n* is 100 in Population 1, and 1000 in the larger Population 2.

Table 7
Population 1: Relative Efficiency when *n =100*

Strata	Method	Variance	Efficiency
3	Cumroot	2,592.41	1.02
	Geometric	2,659.74	
4	Cumroot	1,664.86	0.81
	Geometric	1,355.36	
5	Cumroot	857.63	1.06
	Geometric	917.17	.

Table 8
Population 2: Relative Efficiency when *n =1000*

Strata	Method	Variance	Efficiency
3	Cumroot	1,938.99	0.98
	Geometric	1,898.30	
4	Cumroot	1008.31	1.04
	Geometric	1050.47	
5	Cumroot	678.53	1.00
	Geometric	677.85	

We see from Tables 7 and 8 that the relative efficiency is near 1 in most cases, indicating that the precision of the new method is not substantially different from that of the cumulative root frequency method. In all except one case the efficiency is within .06 of 1. The exception is in Population 1, with $L = 4$ strata where a gain of nearly 20% is observed for the geometric method of stratification; in this case a sample of size 80 would suffice with the geometric method to obtain the same precision as the cumulative root frequency method with n = 100.

6. Summary

This paper gives a simple algorithm for the construction of stratum boundaries in positively skewed populations. It is based on an observation by Cochran (1961) that the coefficients of variation within strata are equal when the optimum boundaries have been achieved. We have shown that, with positively skewed populations, the stratum breaks may profitably be obtained using the geometric progression. This method is easier to implement than the commonly used cumulative root frequency of Dalenius and Hodges (1957). Comparisons of the two methods, carried out with two positively skewed real accounting populations divided into three, four and five strata, show that the new method is as precise as the cumulative root frequency method in most cases.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Irish Research Council for Science, Engineering and Technology.

We would like to thank the referees for their constructive comments, which helped to improve the paper.

REFERENCES

Cochran, W.G. (1961), "Comparison of Methods for Determining Stratum Boundaries", *Bulletin of the International Statistical Institute*, 32, 2, pp. 345-358.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd Edition, New York: Wiley.

Dalenius, T. (1950), "The Problem of Optimum Stratification", *Skandinavisk Aktuarietidskrift*, pp. 203-213.

Dalenius, T and J.L. Hodges (1957), "The Choice of Stratification Points", *Skandinavisk Aktuarietidskrift*, pp. 198-203.

——————— (1959), "Minimum Variance Stratification", *Journal of the American Statistical Association*, 54, pp. 88-101.

Gunning, P. and J. M. Horgan (2004), "A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations", *Survey Methodology*, 2, p. 30.

Hedlin, D. (2000), "A Procedure for Stratification by an Extended Ekman Rule", *Journal of Official Statistics*, 16, pp. 15-29.

Horgan, J.M. (2003), "A List Sequential Sampling Scheme with Applications in Financial Auditing", *IMA Journal of Management Mathematics*, 14, pp. 1-18.

Lavallée, P and M.A. Hidiroglou (1988), "On the Stratification of Skewed Populations", *Survey Methodology*, 14, pp. 33-43.